Version with appendix of paper presented
at ACM Sigmetrics '09 (Best paper award)

The Age of Gossip: Spatial Mean Field Regime

Augustin Chaintreau
Thomson
Paris, France
augustin.chaintreau
@thomson.net

Jean-Yves Le Boudec
EPFL, IC-LCA2
Lausanne, Switzerland
jean-yves.leboudec
@epfl.ch

Nikodin Ristanovic
EPFL, IC-LCA2
Lausanne, Switzerland
nikodin.ristanovic
@epfl.ch

ABSTRACT

Disseminating a piece of information, or updates for a piece of information, has been shown to benefit greatly from simple randomized procedures, sometimes referred to as gossiping, or epidemic algorithms. Similarly, in a network where mobile nodes occasionally receive updated content from a base station, gossiping using opportunistic contacts allows for recent updates to be efficiently maintained, for a large number of nodes. In this case, however, gossiping depends on node mobility. For this reason, we introduce a new gossip model, with mobile nodes moving between different *classes* that can represent locations or states, which determine gossiping behavior of the nodes. Here we prove that, when the number of mobile nodes becomes large, the age of the latest updates received by mobile nodes approaches a deterministic mean-field regime. More precisely, we show that the occupancy measure of the process constructed, with the ages defined above, converges to a deterministic limit that can be entirely characterized by differential equations. This major simplification allows us to characterize how mobility, source inputs and gossiping influence the age distribution for low and high ages. It also leads to a scalable numerical evaluation of the performance of mobile update systems, which we validate (using a trace of 500 taxicabs) and use to propose infrastructure deployment.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless communications; C.4 [**Performance of Systems**]: Modeling Techniques

General Terms

Performance, Theory

Keywords

Epidemic, Gossip, Mean Field, Infrastructure Deployment, Updates, Dynamic content

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.
Copyright 2009 ACM 978-1-60558-511-6/09/06 ...$5.00.

1. INTRODUCTION

Epidemic algorithms were first introduced to maintain consistency for a distributed database in the face of updates [8], offering a simple randomized alternative to complex deterministic algorithms. More recently, the same epidemic principle was introduced in order to forward information in mobile networks where links between nodes are intermittent [17]. In such a context, a node stores information and takes advantage of opportunistic contacts with other nodes to propagate this piece of content in a delay tolerant manner. Epidemic algorithms are simple and scale well to deliver the same content to a large number of hosts, complementing the usual shortcomings of an infrastructure. With the multiplication of mobile devices equipped with short-range wireless radios, carried by either people or vehicles, it becomes possible to receive updated information in a peer to peer manner. However, as the mobility of nodes is typically limited in speed, and much slower than information propagation over wired or wireless links, the delay for such a mobility assisted scheme remains non-negligible. In addition, as each infrastructure or mobile node has a constant capacity to forward information to its neighbors, this delay could grow significantly with the network size.

We consider here an application based on opportunistic content updates through mobile nodes. We assume that a source constantly updates a file that is of interest for all nodes in the network. This file, with its timestamp, is immediately sent to one or several base stations, which then propagate it to mobile nodes. In addition, whenever an opportunistic contact occurs between two mobile nodes, the one with the most up-to-date copy of the file forwards it to the other, following an epidemic principle. We call *age* of a node the age of the latest copy of the file the node has seen so far. The analysis of this simple mechanism is difficult to characterize for a few reasons. First, the age is the time elapsed since the content held by this node was emitted by a base station. This differs from most epidemiological models, which are interested only in the time elapsed since the last infection of this node. In particular, the quantity of interest to us is not the spread of a single packet, but the age distribution across all nodes. Also, contacts between mobile nodes occur in function of their positions rather than in a random uniform way.

We build a model designed to shed light on the above issues. We assume that a mobile node has a *class* attribute that may represent its location, as well as other descriptions of its state. As nodes change positions with time, they switch between classes. Contacts with base stations and contacts between mobile nodes occur at a rate that depends on the current classes of these nodes. We wish to study a very large system, in which the population of nodes follows random mobility between classes, and random opportunistic contacts. Our goal is to characterize the age distribution among the population of nodes for each class.

We make the following contributions.

• We prove that, as the number of nodes participating in the system becomes large (while keeping a constant contact rate per base station/mobile node), this opportunistic content update system converges to a deterministic spatial mean-field regime that is uniquely characterized by a system of partial differential equations (PDEs).

• We show that the system of PDEs can be transformed into a set of ODE problems, and thus numerically easily solved. It may be used as a fast simulation tool, in particular either when traces are not available, or to perform a what-if analysis, or when the number of mobile nodes is very large. In spite of our simplifying assumptions, we demonstrate the accuracy of this tool by comparing with a discrete event simulation using measurement of mobility of 500 taxicabs, collected over 30 days in San Francisco. We also illustrate the use of the tool for infrastructure dimensioning.

• Using the system of PDEs, we show that the mean field regime can be characterized as "infrastructure dominant" or "opportunistic contacts dominant" depending on the contact rates and mobility of the nodes, when the number of classes is small. More generally, for extreme operating points (maintaining a small fraction of nodes with a very up-to-date content, or maintaining the age of content in a high fraction of nodes), we show that the system efficiency is dominated by a single characteristic value.

To the best of our knowledge our work is the first to demonstrate the existence of a spatial mean-field regime for opportunistic content systems, or equivalently, for the age of a gossip in a multi-class system. Its unique feature is to characterize the propagation of information via opportunistic contacts and via mobility among different classes. We believe that the system of equations defining this regime could lead to further simplification and characterization of the multiple dimensions of these systems in the future.

The paper is organized as follows. After presenting related work, we describe in §3 the model in more detail. In Section 4 we prove the convergence of the system to a mean-field regime and provide a unique characterization as the solution of a system of PDEs. We derive in §5 the exact solution for the single-class case, and prove more generally asymptotic results on information propagation. In §6 we compare the numerical solution of the PDEs with discrete event simulations using real measurement, and we use this solution in §7 to solve an infrastructure deployment problem.

2. RELATED WORK

Gossip protocols or epidemic algorithms were used in the past to maintain mutual consistency among multiple database sites [8]. In addition to being simple and scalable, these procedures were shown to be efficient with respect to their deterministic counterparts and robust in the face of topological changes. Most of the works studying gossip protocols assume that a node is equally likely to contact any other node at each time step. It was recently shown that similar performance can be attained when nodes contact each other, according to some general static graph topology [10].

Because epidemic algorithms usually assume that nodes collaborate in an uncoordinated manner, they have also been proposed for routing in ad-hoc or delay-tolerant networks where topology is not known *a priori* [17]. Most of the routing protocols proposed in opportunistic delay-tolerant networks rely on epidemic algorithms as a primitive (usually flooding), which is then further improved using additional information and heuristics to decide which packet to transmit (see, *e.g.,* [5]). The main difference with the previous works mentioned above is that messages between nodes are not exchanged randomly or in a static set of neighbors, but they rather follow contacts created by node mobility.

Close to our work are [12] and [2], which study different epidemic strategies for updates between nodes that are intermittently connected. In [12], a different model of contacts between the nodes is considered, conditions for scalability and optimality of updates' distribution are established. [2] focuses on optimal control. However there are important differences: first, we assume a more general model, where nodes move between classes and contact each other and a base node with rates that depend on the class. Though our model does not include cost, it allows to truly study the influence of mobility and geographical constraint on the performance of epidemic algorithm. Second, we prove convergence to a mean field regime, whereas [2] mentions it as a plausible assumption. Third, we completely characterize the mean field regime by PDEs, which allows us to both obtain efficient solution methods and derive analytical conclusions. In particular, we show that the dynamics of this system follow linear multidimensional ODEs when focusing on low and high age, which gives new insight into the impact of base stations and opportunistic node contacts.

Also close to our work is [3], which compares delivery latency of mesh, base stations and mobile relays in opportunistic systems. Similarities are in the use of a multiclass model for spatial aspects, and in the use of differential equations. Our results are significantly different, though. First, because we focus on opportunistic content updates, we want to characterize the distribution of latency among nodes and classes, rather than the dissemination of a single content. To put it differently, and leaving aside the class attribute of a node, in [3] and [18], the state of one node is a single bit (infected or not) whereas in our case it is a nonnegative real number (the age of the node's content). Note that the age cannot be deduced from the time since the last infection, as it depends on when the content was originally emitted by a base station. Thus we have a new way to evaluate the freshness of disseminated information. Showing convergence to a mean field regime in our case is entirely new (and non trivial), whereas convergence to a mean field regime in the case of one bit of information per node as in [3] follows for example from [13]. Also note that one can derive the extent of infection from the age distribution, so that, in some sense, our model generalizes the model in [3] (but note that [3] focuses on dimensioning rules that are not directly addressed in this paper).

We believe that our characterization of the age of gossip for large systems significantly complements previous works. It can be used as a building block to address future issues of cost efficiency when mobility plays an important role.

3. SPECIFICATIONS AND MODEL

3.1 A Multiclass Approach

We generally assume that nodes are distributed in a finite number of classes. A node may belong to only one class at any given time, although its class may change with time. This allows us to capture the mobility of nodes, as different classes can represent different locations or cells. Note that classes can capture other aspects of the system (type of nodes, wireless interface, etc.). We assume that nodes in the same class are statistically equivalent. Essentially it means that two different nodes, under the same condition of position, type, interface (captured via their class), behave statistically the same with respect to the evolution of their current information age.

Assumptions.

A collection of N nodes move and collect updates according to the following three dynamics.

• **Mobility:** There exists a finite collection $\mathcal{C} \subset \mathbb{N}$ of C classes, and each node belongs to a only one class at any given time. We call $\rho_{c,c'}$ the rate of movements from class c to class c' per time unit.

• **Source Emission:** At any time a node can receive updated information directly from the source (*i.e.* a server maintained in an infrastructure). This may happen with rate μ_c for nodes that are in class c.

• **Opportunistic Contacts:** A node may meet opportunistically with nodes in the same or other classes. In this case we assume that the node with the most recent information transmits it to the other. We define the parameters η_c, $c \in \mathcal{C}$ such that, whenever a pair of nodes both are in class c, they meet with a rate $\left(\frac{2\eta_c}{N-1}\right)$. This implies that the total contact rate in one class is $\frac{N_c(t)(N_c(t)-1)}{N-1}\eta_c$ where $N_c(t)$ is the number of nodes currently in class c.

We also allow for opportunistic contacts among nodes in different classes. This applies to cases where classes represent different types of nodes in the same location, or, as in §6, to contacts across class boundaries, when two classes are mapped to neighboring subareas. We define $\beta_{\{c,c'\}}$, for $c \neq c'$ such that two nodes belonging to classes $c \neq c'$ meet with a rate $\left(\frac{2\beta_{\{c,c'\}}}{N-1}\right)$.

Note that a class may have no infrastructure (*i.e.* $\mu_c = 0$). In this case, the update of information can only come from nodes that visited different classes. It may also represent an inactive state (*i.e.* $\mu_c = \eta_c = 0$) where no meeting occurs.

Example 1 (Homogeneous Network). There is $C = 1$ class. This is the simplest but probably not a realistic model. All nodes are the same and are equally likely to meet with the information source at any time (with rate μ), as well as with each other (with rate η).

Example 2 (Classes as geographical regions). We can map a realistic scenario to classes as follows. We map a geographical area of interest to classes, where each class represents a sub-area. In some classes there is one or more sources of information, in which case μ_c is the aggregate rate of injection of new information at the sources (we explain in §6 how to measure μ_c). In other classes, there is no source, in which case $\mu_c = 0$. We also consider an extra class (class 16 in §6) to represent mobile nodes that are not currently in the area of interest.

We show in §6 that classes do matter, in the sense that a model with just one or two classes gives a poor fit to trace results, whereas one with more classes gives a good fit.

Note that we assume in the model that the total number of nodes N is constant, but, as illustrated in Example 2 above, we can account for a variable number of nodes by introducing an extra class, to represent mobiles that are not present in the area of interest. Thus, with our model, N is in fact an upper bound on the number of nodes in the area of interest. We leave it to a forthcoming paper to extend the model to a variable, possibly unbounded, N.

Metric.

We are interested in the age distribution at any time and in any class. We are interested in the following quantities.

• $u_c^N(t)$ is the fraction of nodes in class c at time t.

• $F_c^N(z,t)$ is the fraction of nodes at time t that are in class c and whose information (obtained from the source or by gossiping) has age $\leq z$. Note that we have, for any $t \geq 0$, $0 \leq F_c^N(z,t) \leq u_c^N(t)$, and $F_c^N(0,t) = 0$, $F_c^N(\infty,t) = u_c^N(t)$.

3.2 Model

The evolution of the system above is captured in continuous time via a drift and jump process. The state of the system at time t is $(\vec{X}^N(t), \vec{c}^N(t)) = \left((X_n^N(t))_{n=1}^N, (c_n^N(t))_{n=1}^N\right)$, with:

$X_n^N(t)$: age of the most recent information held by node n.
$c_n^N(t)$: current class of node n.

The dynamics of ages is essentially characterized by:

• If nodes m, n meet at time t then $X_m^N(t) := X_n^N(t) := \min(X_m^N(t^-), X_n^N(t^-))$.

• If node m meets a base station at time t then $X_m^N(t) = 0$.

• The age of a node increases at rate 1 in an interval where this node does not meet any other nodes nor base stations.

We now formally describes all details of our model.

3.2.1 Evolution of Nodes Between Classes

Let $\{\ K_{n,c,c'} \mid n \in N, c \in \mathcal{C}, c' \in \mathcal{C}, c \neq c'\ \}$ be $N \times C \times (C-1)$ independent Poisson Processes such that $K_{n,c,c'}$ has a rate $\rho_{c,c'}$. Each point of this process denotes a possible transition from the class c to the class c' for node n (the transition always exist, but it has no effect unless the node n is currently in state c). Thus

$$d\vec{c}^N = \sum_{n \in N} \sum_{c \in \mathcal{C}, c' \in \mathcal{C}, c' \neq c} (c' - c) \cdot 1_{\{c_n^N = c\}} \cdot \vec{e}_n dK_{n,c,c'} \ ,$$

where $\vec{e}_m$ is the $N \times 1$ vector with 0 at all components except the mth which is equal to 1. We can rewrite the fraction of nodes in class c, u_c for any N and any time t as:

$$u_c^N(t) = \frac{1}{N} \sum_{n=1}^{N} 1_{\{c_n^N = c\}} \, .$$

The process $\left\{\ (u_c^N(t))_{c \in \mathcal{C}} \ \middle|\ t \geq 0\ \right\}$ may also be thought of as the occupancy measure of the vector $\vec{c}^N$ with values in $\mathcal{C}$. In other words, it characterizes the values taken by all the coordinates of $\vec{c}^N$ but ignores to which coordinates each value corresponds.

If we assume that the process above satisfies the initial conditions that converge to a deterministic limit $(d_c)_{c\in\mathcal{C}}$:

$$\forall c \in \mathcal{C}\,,\ \lim_{N\to\infty} u_c^N(0) = d_c\,,\ \left(\text{ for } \ d_c \geq 0, \sum_{c\in\mathcal{C}} d_c = 1\right) \quad (1)$$

then as N becomes large, Kurtz's theorem (see *e.g.,* [13]) states that the process of $\{\ (u_c^N(t))_{c\in\mathcal{C}}\ |\ t\geq 0\ \}$ converges to a deterministic limit $\{\ (u_c(t))_{c\in\mathcal{C}}\ |\ t\geq 0\ \}$ which is the unique solution of the following ODE problem:

$$\begin{cases} \forall c\in\mathcal{C}\,, & \frac{\partial u_c}{\partial t} = \sum_{c'\neq c}\rho_{c',c}u_{c'} - \left(\sum_{c'\neq c}\rho_{c,c'}\right)u_c \\ \forall c\in\mathcal{C}\,, & u_c(0) = d_c\,. \end{cases} \quad (2)$$

By Cauchy-Lipschitz theorem, for any boundary condition $(d_c)_{c\in}$ this ODE problem admits a unique solution. Following classical notation, we denote the value at time t of the solution for boundary condition d by $u_c(t|d)$.

Assuming that the matrix ρ is irreducible, we may consider the stable mobility regime where $u_c(t) = \tilde{u}_c$ independently of t and is defined as the unique solution of

$$\forall c\in\mathcal{C}\,,\ \tilde{u}_c(\sum_{c'\neq c}\rho_{c,c'}) = \sum_{c'\neq c}\rho_{c',c}\tilde{u}_{c'} \ \text{ and } \ \sum_{c\in\mathcal{C}}\tilde{u}_c = 1\,. \quad (3)$$

3.2.2 Propagation of Information

Let $A_{n,c}$, $n\in N, c\in\mathcal{C}$ be $N\times C$ independent Poisson Processes such that $A_{n,c}$ has a rate μ_c. Each point of this process denotes possible information received by n directly from the source in class c (the transition always exist, but it has no effect unless the node n is currently in class c).

Let $B_{m,n,c}$ $m\in N, n\in N, m<n, c\in\mathcal{C}$ be $\frac{N\times(N-1)}{2}\times C$ independent Poisson Processes such that $B_{m,n,c}$ has a rate $\frac{2\cdot\eta_c}{N-1}$. Each point of this process denotes a possible opportunistic contacts for the pairs $\{\ m,n\ \}$, occurring in the class c (the transition always exist, but it has no effect unless the nodes n and m are currently both in class c).

Similarly, define $C_{m,n,\{c,c'\}}$ for $m\in N, n\in N, m<n, c\in\mathcal{C}, c'\in\mathcal{C}, c<c'$ be $\frac{N\times(N-1)}{2}\times\frac{C\times(C-1)}{2}$ independent Poisson Processes such that $C_{m,n,c,c'}$ has a rate $\frac{2\cdot\beta_{\{c,c'\}}}{N-1}$. Each point of this process denotes a possible opportunistic contacts for the pairs $\{\ m,n\ \}$, occurring when m or n is in class c and m or n is in class c' (the transition always exist, but it has no impact unless the nodes n and m are currently one in classes c, the other in class c').

$$\begin{aligned} &d\vec{X}^N = \vec{1}dt - \textstyle\sum_{n\in N}\sum_{c\in\mathcal{C}} X_n^N \cdot 1_{\{c_n^N=c\}}\cdot\vec{e}_n dA_{n,c} \\ &+ \sum_{m<n}\sum_{c\in\mathcal{C}}\Big[1_{\{X_m^N<X_n^N\}}\vec{e}_n\left(X_m^N - X_n^N\right) \\ &+ 1_{\{X_m^N>X_n^N\}}\vec{e}_m\left(X_n^N - X_m^N\right)\Big]1_{\{c_n^N=c\}}1_{\{c_m^N=c\}}\cdot dB_{m,n,c} \\ &+ \sum_{m<n}\sum_{c<c'}\Big[1_{\{X_m^N<X_n^N\}}\vec{e}_n\left(X_m^N - X_n^N\right) \\ &+ 1_{\{X_m^N>X_n^N\}}\vec{e}_m\left(X_n^N - X_m^N\right)\Big]1_{\{\{\ c_n^N, c_m^N\ \}=\{\ c,c'\ \}\}}\cdot dC_{m,n,c,c'}\,. \end{aligned}$$

We define the occupancy measure of $\vec{X}^N(t)$ in class c by:

$$M_c^N(t) = \frac{1}{N}\sum_{n=1}^{N} 1_{\{c_n^N(t)=c\}}\delta_{X_n^N(t)}\,.$$

$F_c^N(z,t)$ (*i.e.,* the fraction of nodes that are in class c and with ages under z) is

$$F_c^N(z,t) = M_c^N(t)\left([0;z]\right) = \int_0^z M_c^N(t)(du)\,.$$

4. THE MEAN-FIELD REGIME

4.1 Mean-Field Limit

The main result of this section is that, as shown in §3.2.1 for the evolution of nodes between classes, the evolution of the information age among nodes, when N gets large, becomes close to a deterministic limit characterized by differential equations.

We assume that the initial conditions of the system, as N gets large, converge to a deterministic limit. In other words, the occupancy of classes by nodes converges to a deterministic vector $(d_c)_{c\in\mathcal{C}}$ according to Eq.(1), and the initial occupancy measure $M_c^N(0)$ of ages in each class converges weakly to a deterministic distribution m_c^0, with CDF F_c^0.

THEOREM 4.1. *As N gets large, the collection of occupancy measures M_c^N converges in distribution to deterministic processes $\{\ m_c(t)\ |\ t\geq 0\ \}$.*

If m_c^0 admits a density, then $m_c(t)$ has a density for all t and its CDF $F_c(z,t)$ is the unique solution of the following PDE problem

$$\begin{cases} \forall c\in\mathcal{C}\,, & \dfrac{\partial F_c(z,t)}{\partial t} + \dfrac{\partial F_c(z,t)}{\partial z} = \\ & \displaystyle\sum_{c'\neq c}\rho_{c',c}F_{c'}(z,t) - \left(\sum_{c'\neq c}\rho_{c,c'}\right)F_c(z,t) \\ & +\left(u_c(t|d) - F_c(z,t)\right)\left(2\eta_c F_c(z,t) + \mu_c\right) \\ & \displaystyle +\left(u_c(t|d) - F_c(z,t)\right)\sum_{c'\neq c}2\beta_{\{c,c'\}}F_{c'}(z,t) \\ \forall c\in\mathcal{C}\,, & \forall t\geq 0\,,\ F_c(0,t) = 0 \\ \forall c\in\mathcal{C}\,, & \forall z\geq 0\,,\ F_c(z,0) = F_c^0(z)\,. \end{cases} \quad (4)$$

where $u_c(t|d)$ denotes the solution of Eq.(2).

How to explain the equation above. Though we give a mathematical proof in the next section, it is possible to heuristically derive Theorem 4.1 by considering the mean field limit for the densities. The theorem implies that $F_c(z,t)$ admits a density $f_c(z,t)$ at all times t if it has one at time 0; intuitively, the density should satisfy for all c:

$$\begin{aligned} & f_c(0,t) = \mu_c\cdot u_c(t) \\ & \frac{\partial f_c(z,t)}{\partial t} = -\frac{\partial f_c(z,t)}{\partial z} - \mu_c f_c(z,t) \\ & \quad + \quad \sum_{c'\neq c}\rho_{c',c}f_{c'}(z,t) - \left(\sum_{c'\neq c}\rho_{c,c'}\right)f_c(z,t) \\ & \quad + \quad 2\eta_c\left[(+1)\times(u_c(t) - F_c(z,t))\cdot f_c(z,t)\right. \\ & \qquad\qquad \left. + (-1)\times f_c(z,t)\cdot F_c(z,t)\right] \\ & \quad + \quad \sum_{c'\neq c}2\beta_{\{c,c'\}}\left[(+1)\times(u_c(t) - F_c(z,t))\cdot f_{c'}(z,t) + \right. \\ & \qquad\qquad \left. + (-1)\times f_c(z,t)\cdot F_{c'}(z,t)\right]\,. \end{aligned}$$

The second equation can be interpreted using the different possible transitions from the point of view of the current population of nodes in class c and with ages around z: The first term denotes the passage of time. The second term denotes the population removed from age z by source injection. The second line denotes the movement of the node population with age z among different classes. The third and fourth line denotes the impact of opportunistic contacts within the same class and among different classes.

The first transition corresponds to a new node in class c becoming of age z (which is why it is multiplied by $+1$). The node should already be in the class c and the nodes it met should have age z, hence the rate of such a transition is $(u_c(t) - F_c(z,t)) \cdot f_c(z,t) \cdot 2\eta_c$, or if the contacts is among different classes $(u_c(t) - F_c(z,t)) \cdot f_{c'}(z,t) \cdot 2\beta_{\{c,c'\}}$. Last, we have to account for transition where one node in class c is not any more with age z (which explains the -1 for the population) because its age decreases. This node leaving should be of age z, and it should meet a node with an age at most z, hence the transition occurs with rate $f_c(z,t) \cdot F_c(z,t) \cdot 2\eta_c$, or respectively with rate $f_c(z,t) \cdot F_{c'}(z,t) \cdot 2\beta_{\{c,c'\}}$ if this is an opportunistic contacts with another class c'.

The above system of equations may be simplified if we write by convention, when $c = c'$, $\beta_{\{c,c'\}} = \eta_c$. We can then write, as an example, $\sum_{c \in \mathcal{C}} \beta_{\{c,c'\}} = \sum_{c \neq c'} \beta_{\{c,c'\}} + \eta_c$.

$$
\begin{aligned}
& f_c(0,t) = \mu_c \cdot u_c(t) \\
& \frac{\partial f_c(z,t)}{\partial t} = -\frac{\partial f_c(z,t)}{\partial z} - \mu_c f_c(z,t) \\
& \quad + \sum_{c' \neq c} \rho_{c',c} f_{c'}(z,t) - \left(\sum_{c' \neq c} \rho_{c,c'} \right) f_c(z,t) \\
& \quad + \sum_{c'} 2\beta_{\{c,c'\}} \left[(u_c(t) - F_c(z,t)) f_{c'}(z,t) - f_c(z,t) F_{c'}(z,t) \right] .
\end{aligned}
$$

Note that $z \mapsto (u_c(t) - F_c(z,t)) \cdot F_{c'}(z,t)$ is a primitive with regard to z of the terms in the last sum. Therefore, after integrating with regard to z, we obtain Eq.(4).

4.2 Solution of the PDE problem

The PDE problem described by Eq.(4) admits a unique solution, obtained as the transform of a function defined by an ODE problem.

THEOREM 4.2. *The problem in Eq.(4) has a unique solution F given by:*

$$
\forall c \in \mathcal{C}\,,\ F_c(z,t) = \begin{cases} h_c(z|0, u(t-z|d)) & \text{for } z \leq t \\ h_c(t|F_c^0(z-t), d) & \text{for } z > t \end{cases} \tag{5}
$$

where $h(.|b,d)$ denotes the solution of the following ODE problem defined for a function $H : [0;\infty[\to [0;1]^C$:

$$
\begin{cases}
\forall c \in \mathcal{C}\,, & \dfrac{d\, H_c(x)}{dx} = \displaystyle\sum_{c' \neq c} \rho_{c',c} H_{c'}(x) - (\sum_{c' \neq c} \rho_{c,c'}) H_c(x) \\
& \quad + (u_c(x|d) - H_c(x)) \left(\mu_c + 2\eta_c H_c(x) \right) \\
& \quad + (u_c(x|d) - H_c(x)) \left(\displaystyle\sum_{c' \neq c} 2\beta_{\{c,c'\}} H_{c'}(x) \right) \\
\forall c \in \mathcal{C}\,, & H_c(0) = b_c\,.
\end{cases}
$$

In the special case where the class occupancy starts in steady state (i.e. $u(0) = \tilde{u}$), we have $F(z,t) = h(z|0,\tilde{u})$ for $z \leq t$ and thus $F(z,t)$ does not depend on t for $z \leq t$; however, it still depends on t for $z > t$.

4.3 Proof of the Mean Field Regime

We do the proof of Theorem 4.1 for $C = 1$, as the general proof is similar, but the notation becomes much heavier. Due to space limitations, we give only the essential elements of the proof. The proof is based on the method of [13, 15, 7, 4]. The proofs of lemmas can be found in [6].

We use the notation: $\mathcal{P}(\mathbb{R}^+)$ is the set of probabilities on $\mathbb{R}^+$. For $\nu \in \mathcal{P}(\mathbb{R}^+)$ and f bounded defined on $\mathbb{R}^+$, $< f, \nu > = \int_0^{+\infty} f(x) d\nu(x)$. Note that $< f, M^N(t) > = \frac{1}{N} \sum_{n=1}^N f(X_n^N(t))$. The dirac mass at $x \in \mathbb{R}$ is δ_x, so $M^N(t) = \frac{1}{N} \sum_{n=1}^N \delta_{X_n^N}(t)$.

We consider $\left(M^N(t)\right)_{t \geq 0}$ as a stochastic process with samples paths in $D_{\mathcal{P}(\mathbb{R}^+)}[0,\infty)$ (the set of càdlàg functions of $t \in \mathbb{R}^+$, with values in $\mathcal{P}(\mathbb{R}^+)$)[1]. Theorem 4.1 states that M^N converges in distribution to m where m is deterministic, and its set of marginal CDFs satisfy the PDE in Eq.(4).

The main steps are as follows. First, we show that the original system satisfies a martingale problem, obtained by computing the generator and applying it to linear and quadratic functions. Application to linear functions captures the deterministic component of $M^N(t)$, whereas application to quadratic functions captures the variance of $M^N(t)$. More precisely, let $\mathcal{G}^N$ be the generator of M^N; for ϕ a function from $\mathcal{P}(\mathbb{R}^+)$ to $\mathbb{R}$ and $\nu \in \mathcal{P}(\mathbb{R}^+)$:

$$
(\mathcal{G}^N \phi)(\nu) := \lim_{dt \to 0} \frac{1}{dt} \mathbb{E} \left(\phi(M^N(t+dt) - \phi(\nu)) \Big| \, M^N(t) = \nu \right) .
$$

We compute $\mathcal{G}^N$ for ϕ of the form $\phi(\nu) = < h, \nu >$ and $\phi(\nu) = < h, \nu >^2$, where h is an arbitrary test function from $\mathbb{R}^+$ to $\mathbb{R}$ with bounded and continuous first and second derivatives. We show in appendix that:

$$
\mathcal{G}^N(< h, . >)(\nu) = A(h,\nu) + B(h,\nu) + \frac{N}{N-1} C(h,\nu) \tag{6}
$$

with

$$
\begin{aligned}
A(h,\nu) &= < h', \nu > \\
B(h,\nu) &= \mu \left(h(0) - < h, \nu > \right) \\
C(h,\nu) &= 2\eta \int \int 1_{\{x<y\}} \left(h(x) - h(y) \right) d\nu(x) d\nu(y)\,.
\end{aligned}
$$

The term A accounts for aging, B for direct dissemination by the source and C for dissemination by meetings. Further:

$$
\begin{aligned}
\mathcal{G}^N(< h, . >^2)(\nu) &= 2 < h, \nu > (\mathcal{G}^N < h, . > \nu) \\
&+ \frac{1}{N} D(h,\nu) + \frac{1}{N-1} E(h,\nu)
\end{aligned} \tag{7}
$$

with $(\mathcal{G}^N < h, . > \nu)$ given by Equation (6) and

$$
\begin{aligned}
D(h,\nu) &= \mu < (h(0) - h)^2, \nu > \\
E(h,\nu) &= 2\eta \int \int (h(x) - h(y))^2 1_{\{x<y\}} d\nu(x) d\nu(y)\,.
\end{aligned}
$$

The terms D and E represent, in a sense, the variance of $M^N(t)$. As $N \to \infty$, these terms disappear (due to the $1/N$ or $1/(N-1)$ factors) and, informally speaking, this is why the large N limit is non-random. It follows from standard results on jump and drift processes [13] that

LEMMA 4.1. *M^N is a solution to the martingale problem defined by:*

For all functions $h : \mathbb{R}^+ \to \mathbb{R}$ with continuous and bounded first and second derivatives:

$< h, M^N(t) > - < h, M^N(0) > - \int_0^t \mathcal{G}^N < h, . > (M^N(s)) ds$

and

$< h, M^N(t) >^2 - < h, M^N(0) >^2 - \int_0^t \mathcal{G}^N(< h, . >^2)(M^N(s)) ds$

are $M^N(t)-$martingales.

Second, using dominated convergence arguments, we show that any weak limit of a subsequence M^{N_k}, $k \to \infty$ must satisfy the following martingale problem:

[1] It should be possible to extend the proof to a stronger result, namely the convergence in $\mathcal{P}\left(D_{\mathbb{R}+}[0,\infty)\right)$ of $\mu^N = \frac{1}{N} \sum_n \delta_{X_n^N}$, of which M^N is the set of marginals.

DEFINITION 4.1. *PROBLEM MP : For all functions $h : \mathbb{R}^+ \to \mathbb{R}$ with continuous and bounded first and second derivatives:*
$< h, M(t) > - \int_0^t G(h, M(s))ds$
and
$< h, M(t) >^2 - < h, M(0) >^2 - \int_0^t 2 < h, M(s) > G(h, M(s))ds$
are $M(t)-$martingales, where for all $\nu \in \mathcal{P}(\mathbb{R}^+)$:

$$G(h, \nu) = A(h, \nu) + B(h, \nu) + C(h, \nu) \qquad (8)$$

with A, B and C defined in Eq (6).

Third, we show that MP has at most one solution with continuous sample paths and we establish, by means of the next three lemmas, equivalence with the PDE problem:

LEMMA 4.2. *Let $m(t)$ be a solution of the martingale problem MP, with almost all sample paths continuous (in $\mathcal{P}(\mathbb{R})^+$ with Prohorov's topology), and such that $m(0) = m_0$ almost surely, where $m_0 \in \mathcal{P}(\mathbb{R}^+)$. Then almost every sample path of $m(t)$ satisfies: for each function h with continuous derivative:*

$$< h, m(t) > = < h, m_0 > + \int_0^t G(h, m(s))ds\,. \qquad (9)$$

LEMMA 4.3. *The problem in Equation (9), where $m(t)$ is unknown but deterministic and $m_0 \in \mathcal{P}(\mathbb{R}^+)$ is given, has at most one solution.*

LEMMA 4.4. *Assume $F(z, 0)$ is differentiable and let $F(z, t)$ be the unique solution to the PDEs in Eq (4). For every t let $m(t)$ be the measure with CDF $F(., t)$. Then $m(t)$ satisfies Equation (9).*

Fourth, we show that:

LEMMA 4.5. *Any weak limit of a subsequence M^{N_k}, $k \to \infty$ a.s. has continuous sample paths.*

At this point, it follows that any weak limit of a subsequence M^{N_k}, $k \to \infty$ is uniquely determined, thus all that remains to be shown is that the sequence M^N is relatively compact, which follows from:

LEMMA 4.6. *The sequence M^N is tight.*

5. ANALYTICAL RESULTS

5.1 Single-class

We prove in [6] that, for the single class case, the distribution of age seen in a mean field regime has the following explicit expression:

$$F(z,t) = \begin{cases} 1 - \frac{2\eta+\mu}{2\eta+\mu e^{(\mu+2\eta)z}}\,, & \text{when } z \le t \\ 1 - \frac{2\eta+\mu}{2\eta+\frac{2\eta F(z-t,0)+\mu}{1-F(z-t,0)}e^{(\mu+2\eta)t}}\,, & \text{when } z > t \end{cases} \qquad (10)$$

where $F(z, 0)$ is the cdf at time $t = 0$.

This solution for F can be well understood. When z is less than t, gossips for this age are created after initialization, thus they follow the steady state behavior, and the density function around z does not depend on t. Otherwise, when $z > t$ this part of the distribution denotes gossip that was present during the initialization (at time 0 they had the age $(z - t)$). The current value of $F(z, t)$ then depends on the value $F(z - t, 0)$ (*i.e.*, at time 0), and the elapsed time.

Introducing $A = (\mu + 2\eta)$ and $\alpha = \frac{2\eta}{\mu+2\eta}$, Eq.(10) implies

$$f(z,t) = \begin{cases} \frac{(1-\alpha)Ae^{Az}}{\left(\alpha+(1-\alpha)e^{Az}\right)^2}\,, & \text{when } z \le t \\ \frac{e^{At}f(z-t,0)/(1-F(z-t,0))^2}{\left(\alpha(1-e^{At})+\frac{1}{1-F(z-t,0)}e^{At}\right)^2}\,, & \text{when } z > t \end{cases}$$

$$\text{when } z \le t, \quad \frac{\partial f(z,t)}{\partial z} = \frac{(1-\alpha)A^2e^{Az}\left(\alpha - (1-\alpha)e^{Az}\right)}{\left(\alpha+(1-\alpha)e^{Az}\right)^3}\,.$$

As shown in Figure 1, the density may then exhibit two different behaviors depending on the value of α, i.e. of η/μ:

- When $2\eta \le \mu$, the density $f(z, t)$ for $z \le t$ sufficiently large, is always non-increasing in z. In that case, the highest density remains equal to μ around the value $z = 0$, exactly as if no opportunistic contacts occur between the mobile nodes. In other words, for these values of η, gossip through opportunistic contacts does not modify qualitatively the distribution of ages.
- When $\mu < 2\eta$, the density $f(z, t)$ for $z \le t$ admits a maximum at $z_0 = \frac{1}{\mu+2\eta}\ln\left(\frac{2\eta}{\mu}\right)$ with $f(z_0, t) = \mu\left(1 - \frac{(\mu-2\eta)^2}{(\mu+2\eta)^2}\right)^{-1}$. This indicates that gossiping among mobile nodes creates in this case a new typical age of the information seen by a node.

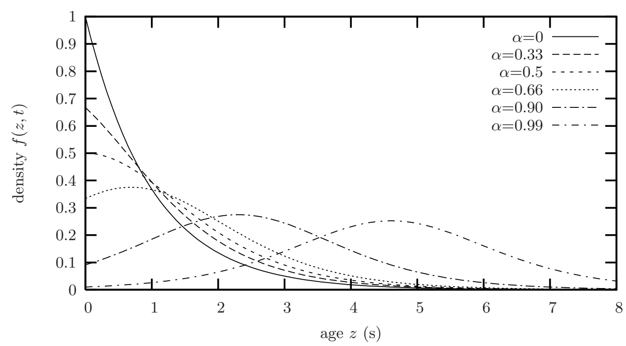


Figure 1: Density of age in a single class for different values of $\alpha = 2\eta/(\mu + 2\eta)$ (normalized by assuming $A = \mu + 2\eta = 1$). A pole exists if and only if $\alpha > 1/2$.

5.2 Multi-Class Asymptotics

The previous section has shown that the single class case can be solved in a closed form. In contrast, the multi-class model does in general not follow a simple expression. This is because the ODE defining h, and therefore F, is multidimensional and not linear. However, as we show below, this ODE may be approximated for low age asymptotics and high age asymptotics by multi-dimensional linear ODEs.

To simplify the analysis, we assume that the class occupation process $u_c(t)$ starts in its steady state $\tilde{u}_c$. In that case, we know that for t sufficiently large $F(z, t) = h(z|0, \tilde{u})$, where h is the solution of the following ODE on H:

$$\begin{cases} \forall c \in \mathcal{C}\,, \ \frac{d\,H_c(x)}{dx} = \sum_{c' \neq c} \rho_{c',c} H_{c'}(x) - \left(\sum_{c' \neq c} \rho_{c,c'}\right) H_c(x) \\ \qquad\qquad + (\tilde{u}_c - H_c(x))\left(\mu_c + \sum_{c' \in \mathcal{C}} 2\beta_{\{c,c'\}} H_{c'}(x)\right) \\ \forall c \in \mathcal{C}\,, \quad H_c(0) = 0 \end{cases} \qquad (11)$$

where we have by convention when $c' = c$, $\beta_{\{c,c'\}} = \eta_c$.

5.2.1 *Low Age*

Note first that, according to Eq.(11) above, the stationary distribution of age $F_c(z)$ satisfies $\forall c \in \mathcal{C}$, $F_c(0) = 0$, as well as the following ODE:

$$\frac{d\,F_c(z)}{dz} = \tilde{u}_c\mu_c + \left(\tilde{u}_c 2\eta_c - \mu_c - \sum_{c'\neq c}\rho_{c,c'}\right)F_c(z)$$
$$+ \sum_{c'\neq c}\left(\rho_{c',c} + \tilde{u}_c 2\beta_{\{c,c'\}}\right)F_{c'}(z) - \sum_{c'\in\mathcal{C}} 2\beta_{\{c,c'\}}F_c(z)\cdot F_{c'}(z)\,.$$

For low value of the age z, for any class c, $F_c(z)$ becomes close to zero. Hence, in the equation above, the last term (that corresponds to a square product of F_c and $F_{c'}$) will be asymptotically negligible when compared to the other two. This equation then becomes a linear differential equation on the vector F (with dimension $C \times 1$), which may be written as $F' = FA + B$ with

$$\begin{cases} A_{c,c} = \tilde{u}_c 2\eta_c - \mu_c - \sum_{c'\neq c}\rho_{c,c'}\,, & \\ A_{c,c'} = \rho_{c,c'} + \tilde{u}_{c'}2\beta_{\{c,c'\}} & \text{if } \ c \neq c'\,, \\ B = \left(\ \mu_0\tilde{u}_0\,,\quad \ldots\,,\quad \mu_C\tilde{u}_C\ \right)\,. & \end{cases}$$

This linear differential equation has an explicit solution in matrix form, given by $F(z) = B\int_0^z e^{(z-s)A}ds$. Using the Taylor series expansion around $z = 0$ it follows that

$$F_c(z) = z\cdot B_c + \frac{z^2}{2}\sum_{c'\in\mathcal{C}} B_{c'}A_{c',c} + o(z^2)\,.$$

Hence for a class $c \in \mathcal{C}$ and low age z, the density of nodes with age z is $\mu_c\tilde{u}_c$, and its derivative at $z = 0$ is given by

$$\mu_c\tilde{u}_c(\tilde{u}_c 2\eta_c - \mu_c - \sum_{c'\neq c}\rho_{c,c'}) + \sum_{c'\neq c}\mu_{c'}\tilde{u}_{c'}\left(\rho_{c',c} + \tilde{u}_c 2\beta_{\{c',c\}}\right)\,.$$

If we assume that contacts do not occur between different classes (*i.e.,* $\beta_{c,c'} = 0$ for all $c' \neq c$), we deduce.

$$\frac{df_c(z)}{dz} = \mu_c\tilde{u}_c\left(\tilde{u}_c 2\eta_c - \mu_c\right) + \sum_{c'\neq c}\left(\mu_{c'} - \mu_c\right)\tilde{u}_{c'}\rho_{c',c}\,. \quad (12)$$

Eq.(12) can serve to classify classes: We may say that a class has “dominant opportunistic contacts” if the density is initially increasing. Such a situation indicates that there exists a pole, or maximum of density, for nodes with positive ages, which is maintained by opportunistic contacts. In contrast, a class where the density is initially decreasing indicates that, at least in the first order, the age of a node rarely benefits from opportunistic contacts. The age in such a class is mostly maintained by infrastructure and we say that the class has “dominant infrastructure”.

One immediately sees from the above equation that

• When infrastructure is deployed uniformly (*i.e.,* μ_c does not depend on c), then a class has dominant infrastructure (resp. dominant opportunistic contacts) if and only if $2\eta_c\tilde{u}_c < \mu_c$ (resp. $2\eta_c\tilde{u}_c > \mu_c$). Note that this characterization does not depend at all on the rates of movement between different classes.

• As expected, any class with no infrastructure coverage (such that $\mu_c = 0$) has dominant opportunistic contacts (this is obvious as the density is initially null).

• If infrastructure is deployed on a subset of non-neighboring classes (for instance, if it is deployed in a single class), then a class with infrastructure has a dominant infrastructure iff

$$2\eta_c\tilde{u}_c < \mu_c + \sum_{c'\neq c}\rho_{c,c'}\,.$$

In other words, for a *sparse* infrastructure (*i.e.,* infrastructure is never found in two neighboring classes), a class is dominated by opportunistic contacts only if the rates of adhoc contacts among nodes compensate both the infrastructure and departure rate from this class.

In a more general case, when infrastructure is deployed in many classes (in different degrees), the characterization above permits the following conclusion: It tells, in the first order that deals with low age, whether a system of mobile nodes significantly benefits from mobility and opportunistic contacts, or if the benefit is only marginal. This conclusion is reached by estimating parameters, and computing the density derivative through Eq.(12).

5.2.2 High Age

Let us denote by $G_c(z)$ the fraction of nodes that are in class c and have an age higher than z, i.e. $G_c(z) = \tilde{u}_c - F_c(z)$. We expect that $G_c(z) \to 0$ as $z \to \infty$. We may change variable F_c in G_c in the associated differential equation above. We find that $G_c(z)$ satisfies

$$\frac{d\,G_c(z)}{dz} = -\left(\sum_{c'\in\mathcal{C}}\tilde{u}_{c'}2\beta_{\{c,c'\}} + \mu_c + \sum_{c'\neq c}\rho_{c,c'}\right)G_c(z)$$
$$+ \sum_{c'\neq c}\rho_{c',c}G_{c'}(z) + \sum_{c'\in\mathcal{C}}2\beta_{\{c,c'\}}G_c(z)\cdot G_{c'}(z)\,.$$

For high value of z we can expect $G_c(z)$ to be small for any class c. Therefore, the last term in the expression above becomes negligible and G follows the multi-dimensional linear equation $G' = GA$, where

$$\begin{cases} A_{c,c} = -\sum_{c'\in\mathcal{C}}\tilde{u}_{c'}2\beta_{\{c,c'\}} - \mu_c - \sum_{c'\neq c}\rho_{c,c'}\,, & \\ A_{c,c'} = \rho_{c,c'} & \text{if } c \neq c'\,. \end{cases}$$

The solution of this equation is $G(z) = G(z_0)e^{(z-z_0)A}$. The matrix A satisfies that all terms outside the diagonal are non-negative, which is sometimes called a Metzler-Leontieff matrix, or ML-matrix (see [16], Chap.2). If we further assume that all classes of nodes communicate, this matrix is irreducible and we can apply Theorems 2.6 and 2.7 in [16] and obtain that A has a dominant eigenvalue τ such that $\tau \in \mathbb{R}$, and $\mathrm{Re}(\lambda) < \tau$ for any eigenvalue λ of A; furthermore, $\tau < 0$.

It follows that there exists a constant K_c such that $G_c(z) = K_c e^{z\tau} + o(e^{z\tau})$. Thus τ characterizes the rate of decay of the age distribution, for all classes. The numerical value of τ may be obtained as the dominant eigenvalue of A. It is contained within the following bounds ([16]):

$$\min_c\Big(-\sum_{c'\in\mathcal{C}}\tilde{u}_c 2\beta_{\{c,c'\}} - \mu_c\Big) \leq \tau \leq \max_c\Big(-\sum_{c'\in\mathcal{C}}\tilde{u}_c 2\beta_{\{c,c'\}} - \mu_c\Big)\,.$$

6. VALIDATION WITH TRACES

In this section we test the accuracy of the model and the mean-field approximation. We use a large, publicly available data set of GPS traces collected by 500 San Francisco's taxi cabs during a period of more than two years [1]. It allows us to generate contact traces between Yellow Cabs (our mobile nodes), and base stations that we can place arbitrarily.

6.1 Validation Setup

In our scenario, the San Francisco Bay Area is divided into 16 classes, as shown in Figure 2. Fifteen classes are obtained using a regular square grid. Each of them corresponds to a region of about 4 sq km. The 16th class surrounds the other classes and contains the area outside classes 1-15. Its existence is important, as it keeps the number of mobile nodes in the system rather constant. Base stations are placed in fixed locations, and we assume that they always have fresh information from a source server. We assume that each mobile node (*i.e.,* a Yellow Cab) is equipped with a short-range radio that allows for the exchange of data upon a meeting with base stations or other mobile nodes. As before, upon a meeting with a base station, a mobile node receives fresh information; and a meeting between two mobile nodes results in both of them having the latest information available in any of them before the meeting.

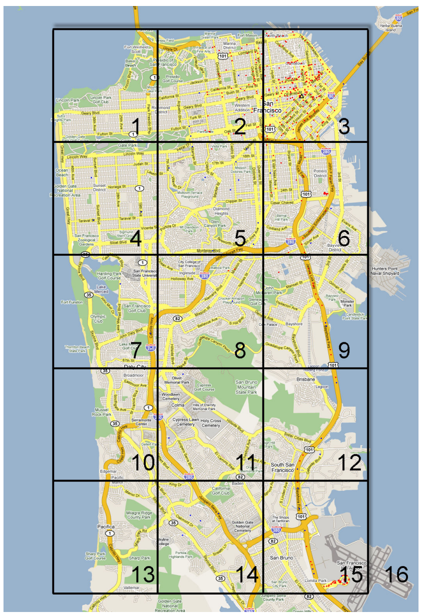


Figure 2: The Bay Area is split into 16 classes.

Data Sets. We use GPS position records, measured approximately once per minute, which were collected within the *Cabspotting* project [1] that aims at visualizing the aspects of everyday life in SF. About 500 Yellow cab vehicles that operate in the area are equipped with GPS receivers. Recorded data is sent to the central dispatcher and stored in the database. Each GPS record contains the cab's ID, current location, as well as the time stamp (as Unix epoch). This allows us to reconstruct the path of each individual mobile node for the past two years.

We consider a 30 day GPS trace, from May 17 to June 15, 2008. We observe the 16 hour periods between 8 a.m. and midnight. We want to avoid night-time, when the number of active cabs drops.

Generation of Contact Traces. In order to obtain an artificial contact trace from an existing GPS trace, we first have to define ranges, for both mobile nodes and base stations. We also have to define the notion of *meeting* between two mobile nodes or a mobile node and a base station.

We assume that mobile nodes' radios, as well as base stations, have a range of $200m$. This corresponds to the envisioned range in vehicular communications [11], and it is a bit longer than the ranges of 802.11 devices ($\sim 140m$) or Bluetooth Class 1 devices ($\sim 100m$). Each mobile node performs scanning once per minute, looking for base stations and other mobile nodes in range. Once another mobile node or a base station is discovered, we use interpolation to make sure that the contact lasts at least 10 seconds. So, we assume that a meeting between two mobile nodes, or a mobile node and a base station happened if, during scanning, a mobile node detected another mobile node, or a base station and their contact lasted for at least 10 seconds. As shown in [5], a real system implemented on buses equipped with 802.11b radios has an average transfer opportunity duration of 10.2s, which is sufficient to exchange on average 1.2MB of data. Contacts between mobile nodes, and with a base station, can occur between scanning periods. We decide to ignore these contacts as most of the existing wireless technologies do not allow scanning and data exchange at the same time.

Using the provided definition of a *meeting*, we run a simulation (written in Java) and obtain the contact trace.

Parameter Settings. The input parameters for the model and the mean field approximation, as defined in Section 3.1, are μ_c, η_c, $\beta_{c,c'}$ and $\rho_{c,c'}$. For each class, we extract them from the contact traces as follows:

$$\mu_c(t) = \frac{N_{c,ub}(t)}{N_c(t)} \ , \ \mu_c = \frac{1}{60}\sum_{t=t_0}^{t_0+60} \mu_c(t)\,,$$

$$\eta_c(t) = \frac{N_{c,uu}(t)}{u_c(t) * (N_c(t) - 1)} \ , \ \eta_c = \frac{1}{60}\sum_{t=t_0}^{t_0+60} \eta_c(t)\,,$$

$$\beta_{c,c'}(t) = \frac{N_{c,c',uu}(t)}{2 * N(t) * u_c(t) * u_{c'}(t)} \ , \beta_{c,c'} = \frac{1}{60}\sum_{t=t_0}^{t_0+60} \beta_{c,c'}(t)\,,$$

$$\rho_{c,c'}(t) = \frac{N_{c,c',trans}(t)}{N_c(t)} \ , \ \rho_{c,c'} = \frac{1}{60}\sum_{t=t_0}^{t_0+60} \rho_{c,c'}(t)\,.$$

where for any time slot t (in minute), $N(t)$ is the total number of nodes; for any c we denote by $N_c(t)$ (resp. $u_c(t)$) the number (resp. the fraction) of nodes in class c, and we denote by $N_{c,ub}(t)$ (resp. $N_{c,uu}(t)$) the number of meeting between mobile nodes and base stations (resp. between two mobile nodes) during the time slot t; finally, for any classes $c \neq c'$, we denote by $N_{c,c',uu}(t)$ (resp. $N_{c,c',trans}(t)$) the number of meetings between nodes of different classes (resp. the number of transitions from c to c') during time slot t. As shown above, per hour values of the parameters are calculated by averaging their per minute (per time slot) values over the period of one hour. The necessary per minute values are extracted from the generated contact trace.

The values of the input parameters indicate that the node distribution is highly skewed: 75% of the nodes are contained within 4 popular classes (2,3,6 and 15, *i.e.,* city center and airport); nodes spend on average 12 to 40 min in one of these classes before moving; 10% of the nodes are contained in surrounding classes (1,4,5,9 and 12) where nodes stay less time (4 to 12mn before moving). Class 16 contains roughly 10% of "persistent" nodes that remain in this class two hours on average. All the other classes contain in total 5% of the nodes; class 13 is generally empty. The rate of meetings for two given nodes within the same class is generally between (1/60mn) and (1/80mn); this contact rate is higher in classes 9,12,15 (1/20mn), and much smaller in 10,11,13,16 (under 1/200mn). Contacts between the nodes in different classes are almost negligible (with rate less than 1/2000mn).

Running the Simulations. The calculated input parameters are used twice: first, to simulate the random model described in Section 3.2 with $N = 500$ nodes, and, second, to evaluate the mean field limit by solving the associated ODEs (Theorem 4.2) in Matlab[2]. The contact trace itself is used directly for an event-driven simulation.

In all three cases, we get the corresponding age distributions for each minute of observation.

6.2 Comparison of Trace, Model & MF Limit

We now compare the age distributions obtained from the trace, the model and the mean-field approximation, for the case of a single base station, placed in class 3. In terms of contacts, a given node in class 3 meets the base station with a rate (1/45mn). Simulations begin at 8 a.m. We set the initial age of information for each mobile node to 8 hours, in line with the night-time inactivity.

Figure 3 shows the trace, the model and the mean-field CDFs for ages in different classes, observed at 1 and 8 p.m. by the three methods mentioned above. CDFs were estimated every minute and averaged over an hour. We omit classes 13 and 16 as they are of less interest and generally not connected to the rest of the network.

The distribution of age, obtained from the model, shows a very good match with the distribution taken from the traces, in particular for popular classes $(2, 3, 6, 15)$ and their surroundings $(1, 4, 5, 6, 12)$. This means that our modeling assumptions succeeded in accurately capturing the collective behavior of gossip among taxis. Some discrepancies are observed for peripheral classes, which may be explained as follows: In classes with very few mobile nodes, the age of a single mobile node (which stopped for some reason too far from the main road and cannot receive an update), can create a significant difference between the trace and the model. Indeed, classes 10, 11 and 14 contain on average 1.1, 2.1 and 2 mobile nodes respectively.

The mean-field limit matches the model extremely well, except, again, for discrepancies observed in peripheral areas.

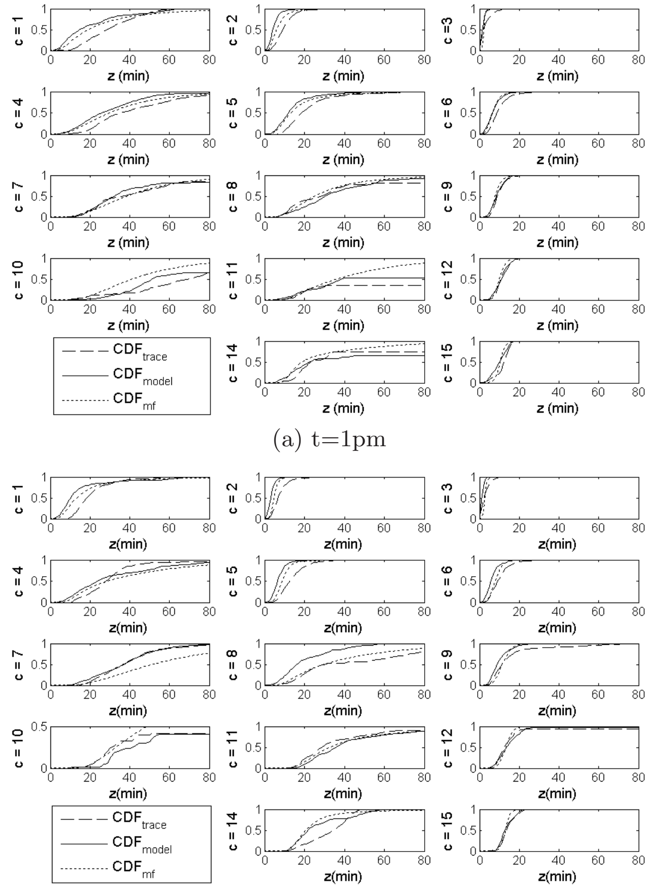
6.2.1 The Importance of Being Opportunistic

For update applications, the quality of service (in terms of age) can be measured as the fraction of mobile nodes in each class whose age is lower than a given threshold. We now compare this metric for the cases when opportunistic exchanges between mobile nodes are allowed, and when the dissemination is performed only through the base station.

Figure 4 shows the percentage of mobile nodes in each class that have age less than 20 minutes at $1p.m.$ and $8p.m.$ (300 and 720 minutes after the start of observation), obtained from the trace, the model and the mean-field approximation. The figure also displays the same percentage for the case where only the base station is allowed to disseminate information; this percentage is generally very small and remains under 20%, even in class 3 where the base station is located. Using opportunistic contacts between mobile nodes, in contrast, allows us to significantly improve percentages in all classes, and to approach very high percentages in classes 2, 3, 6, 9, 12, 15 (which together contain 80% of the nodes).

These results are best interpreted using a spatial representation as in Figure 5, where data provided in the upper panel of Figure 4 are shown spatially for the trace and the mean-field approximation.

[2]The value of $\tilde{u}_c$ in the ODE is obtained from Eq.(3).



(a) t=1pm

(b) t=8pm

Figure 3: CDFs for classes 1-12 and 14-15, obtained from the trace, the model and the mean field limit. The CDFs show the age distribution (z) in different areas of San Francisco at 1 p.m. and 8 p.m., for the case of a single base station, placed in class 3.

We observe that classes benefit differently from the base station located in class 3. First, classes 2 and 6 benefit from it as they are immediate neighbors. Second, classes 9, 12 and 15 benefit from the cars traveling on the highway between the city center and the airport, as well as the high meeting rate between cars in these classes. Classes 1 and 5, although geographically closer to class 3, benefit less from opportunistic contacts, due to bias in the mobility. All other classes benefit only marginally as the density of cars and exchanges are too small.

In summary, the opportunistic contacts are useful as they significantly improve the availability of accurate information in the network and can compensate for a lack of infrastructure. The improvement depends critically on the node density, the mobility and the opportunistic contact rates and it is accurately captured by the mean-field limit.

6.2.2 The Importance of Being Spatial

We now evaluate the influence of the *spatial* approach on the accuracy of the model, by comparing our previous results to a case where all classes $1 - 15$ have been merged into a single one (so that only 2 classes remain, *i.e.,* $C = 2$)

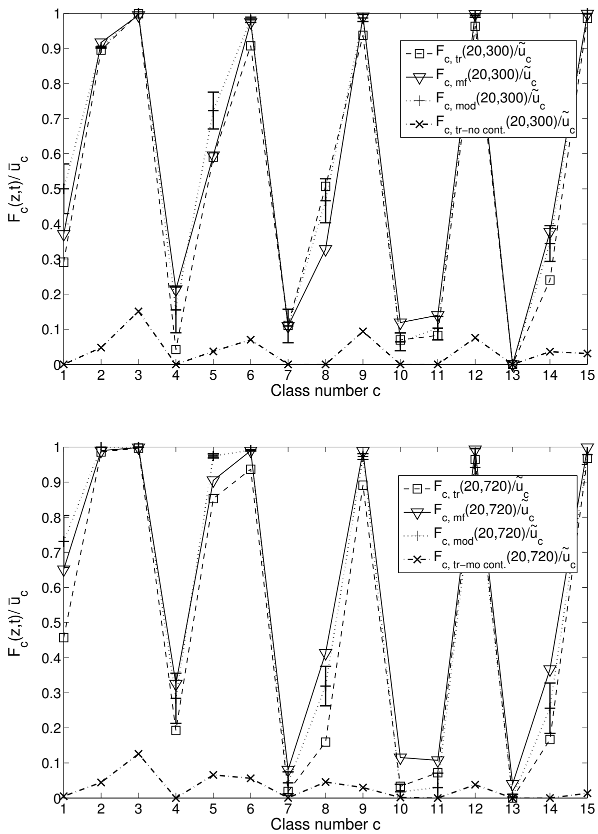


Figure 4: The fraction of mobile nodes in classes 1-15 that have age z<20mn acquired from the trace, the model and the mean field limit, for a single base station placed in class 3. We plot for comparison the values obtained from the trace without opportunistic contacts (bottom curves). Top panel - values at 1 p.m. (t=300mn). Bottom panel - values at 8 p.m. (t=720mn).

Figure 6 displays QQ plots, comparing the age distribution of trace data with the corresponding age distributions obtained from the mean-field approximation, for both of the aforementioned cases ($C = 16$ and $C = 2$). The trace data shown in the figure, for the case $C = 16$, was collected by mobile nodes in classes $1 - 15$, during the afternoon peak hour (5pm-6pm), while the trace data for the 2 class case was collected by mobile nodes in class 1, during the same period. Age samples were taken on a per minute bases. The artificial mean-field data samples were generated from the mean-field CDFs, for the same time interval.

Figure 6(a) suggests that the mean-field and trace age data samples, for the case $C = 16$, come from the same distribution. In contrast, when $C = 2$, (Figure 6(b)), we observe that the mean-field limit underestimates quantiles for low age and almost always overestimates quantiles for high age. This is a clear indication that data came from different distributions.

The results above show that it is essential to capture the diversity of locations (via classes), as they differ radically in terms of expected performance (age distribution). The primary factors are the dependencies between classes created by patterns of mobility (transition matrix $\rho_{c,c'}$) and the contact rates (μ_c, η_c, $\beta_{c,c'}$) that are influenced by mobile node densities and variations in placement of base stations.

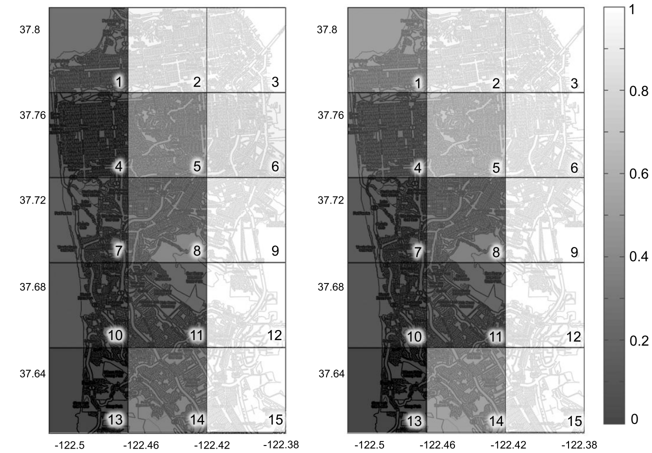
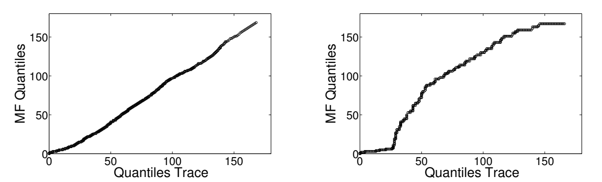


Figure 5: Comparison between the mean field limit and the trace. Percentages of mobile nodes in classes 1-15 with age z<20mn at time t=300mn (1 p.m.).



(a) MF vs Trace (C=16) (b) MF vs Trace (C=2)

Figure 6: The importance of being spatial. QQ plots, comparing the age distributions of trace data and data artificially obtained from the mean field CDF, for 16 class and 2 class scenarios. Time period observed is 5 p.m.-6 p.m.

7. APPLICATION

Let us consider the following problem. We would like to leverage mobility and opportunistic contacts between taxicabs to disseminate news, traffic information or advertising. Each of these applications, however, requires a certain level of infrastructure (base stations). The number and placement of base stations, needed to achieve a given quality of service, are not easy to guess. The answer, in general, depends on the density of nodes in different areas, as well as the transition rates and rates of opportunistic contacts. We demonstrate in this section that a greedy algorithm based on the mean-field limit offers a fast and efficient method for placement of base stations, over multiple classes and a significant improvement over the other simple heuristics.

7.1 Method for Infrastructure Deployment Based on MF Approximation

The problem we solve can be formulated in the following way: For a fixed budget (fixed number of base stations), we would like to find an efficient placement of base stations over a predetermined finite set of classes, based on a range of possible metrics.

Assumptions. We assume that a predefined set of possible locations in each class, where the base stations can be placed, is known to the service planners. Each of these locations carries information about the popularity of the spot. Our assumption is that this piece of information, along with the other input parameters required by the model, can be provided by traffic engineers (traffic counting and estimation models), or based on a trace, collected by some other service in the city (a taxi company for instance).

Metrics. We wish to maximize one of the following objectives ($F_c(z_0, t_0)$ and $\tilde{u}_c$ follow previous definitions):

$metric_1$	$metric_2$	$metric_3$
$\sum_{c=1}^{C'} F_c(z_0, t_0)$	$\sum_{c=1}^{C'} \frac{F_c(z_0,t_0)}{\tilde{u}_c}$	$\min_{c=1,\ldots,C'} \frac{F_c(z_0,t_0)}{\tilde{u}_c}$

Maximizing $metric1$ is a global "per mobile node" objective; it tends to maximize the number of mobile nodes, in all classes, that have an age lower then z_0 during the peak hour (t_0); $Metric2$ is a "per class" metric; using this metric we try to achieve more even distribution of mobile nodes with ages lower than z_0, over the observed 15 classes. Finally, $metric3$ focuses on the class with the "worst" value of the age, and tends to decrease the gap in quality that exists between this class and the other classes; this metric can be used for instance if we want to achieve some minimal QoS in all classes. We denote the total number of classes, where we plan to place base stations, by C'. In our particular case $C' = 15$, as we do not place any base stations in class 16 (our goal is not to maximize the quality in class 16).

Placement of base stations. The algorithm we propose for the placement of base stations is a greedy algorithm (see [9]). Let us denote the total number of base stations by S, and the the number of base stations placed in class c by a_c. The cost will be defined as the total number of base station (*i.e., cost* $= \sum_{c=1}^{C'} a_c$). As explained in *Assumptions*, we assume that the dependency $\mu_c(a_c)$ is known to the service planners, along with the other input parameters for our model. As defined in 3.1, μ_c denotes a contact rate with the base stations, per mobile node, per time unit, inside class c. A base station placed in class c cannot be seen from other classes, but only within a region, limited by the base station's range, inside class c. We start adding base stations one by one. For each base station there are C' possible placement options, one for each of the classes to which base stations are being added. Here we apply the *greedy* approach and use the ODEs (Theorem 4.2) to evaluate which placement, out of C' possibilities, brings the most benefit to the observed metric. The base station is then placed accordingly. The procedure is repeated S times until all S base stations are placed. Algorithm requires the system of ODEs to be solved SxC' times.

Example. Here we provide a numerical example for the placement method described above, based on the taxicab scenario described in Section 6. We assume that 30 base stations should be placed. The input parameters for the system of ODEs are known, as well as the dependency $\mu_c(a_c)$. We used input parameters for the afternoon peak hour ($5p.m.-6p.m.$). The goal is to choose values of a_c for each of the fifteen classes where base stations can be placed.

Class num	Number of base stations per class				
	metr1	*metr2*	*metr3*	*unif.*	*prop.*
$c = 1$	3	2	0	2	0
$c = 2$	0	0	0	2	6
$c = 3$	0	0	0	2	10
$c = 4$	7	5	0	2	0
$c = 5$	0	0	0	2	2
$c = 6$	1	1	0	2	5
$c = 7$	2	2	1	2	0
$c = 8$	3	3	1	2	0
$c = 9$	0	0	1	2	1
$c = 10$	2	3	5	2	0
$c = 11$	1	1	9	2	0
$c = 12$	0	0	0	2	1
$c = 13$	5	7	7	2	0
$c = 14$	6	5	5	2	0
$c = 15$	0	1	1	2	5

Table 1: Placement of 30 **base stations in classes** $1-15$**, acquired from the *greedy placement*, which uses the ODEs for quality estimation, for** 3 **metrics defined in this section. Uniform and proportional placements of base stations are also shown.**

Table 1 shows results obtained with 3 different metrics defined in this section, as well as the values of a_c for proportional and uniform placements. In each of the metrics we use $z_0 = 20min$ as the value for age.

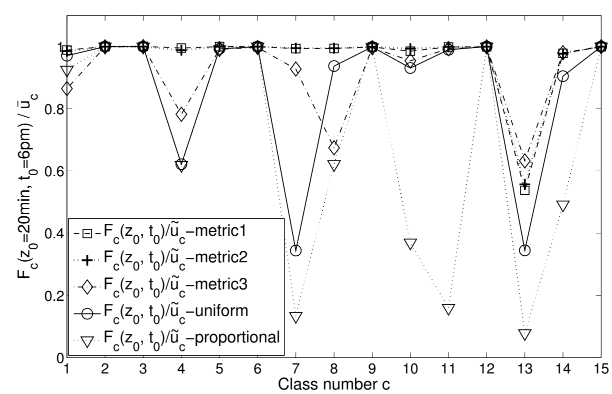
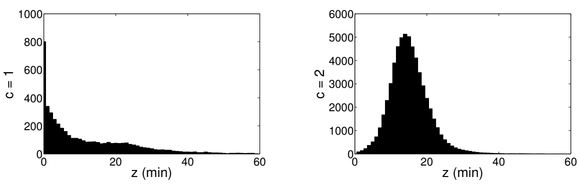


Figure 7: QoS achieved in classes $1-15$ **using *greedy method* for placement of base stations, based on MF approximation. The curves show proportions of mobile nodes in each class with age** $z < 20$ **minutes, at time** $t_0 = 6p.m.$**, for** 3 **metrics, defined earlier in this section, as well as for proportional and uniform placements of base stations.**

Figure 7 shows the effect these placements have on quality of service. It displays the fraction of mobile nodes in each class that have age lower then 20mn; we see that $metric1$ and $metric2$ provide similar results, even though the placements of base stations for these two metrics are different. $Metric3$ sacrifices efficiency for fairness, and degrade performance of most classes, to reach a marginal improvement in the worst case class. Finally we see that proportional placement of base stations, based on the density of mobile node, results in worse performance, than the less sophisticated uniform placement with 2 base stations in each class.

7.2 Classification of Classes

While the tail of our trace is insufficient for an evaluation of high age asymptotic, (Section 5.2.2), low age asymptotic analysis proves to be an efficient tool for classification of classes proposed in Section 5.2.1. For different placements of base stations, evaluated in the previous example, the sign of the first derivative (Eq.(12)) always correctly classifies the classes into those with "dominant opportunistic contacts" and those with "dominant infrastructure", i.e. it always matches the characteristic shapes of the age histograms, for these two types of classes (Figure 8). Due to the lack of space, the complete set of results will be provided in the associated tech-report.



(a) dominant infrastructure (b) dominant opp. contacts

Figure 8: Typical shapes of age histograms for class with "dominant infrastructure" (c=1) and class with "dominant opportunistic contacts" (c=2). base station placement corresponds to $metric1$ (Table 1).

For the case of a base station put in class 3, which corresponds to the scenario with base stations placed in a subset of non-neighboring classes (Section 5.2.1), we conclude that class 3 is also dominated by opportunistic contacts.

8. CONCLUSION

It is sometimes possible to rely on the mobility of nodes to mimic a simple randomized scheme where nodes contact each other at random. However, as we have shown in a simple case (maintaining recent updates of a content with opportunistic contacts), the mobility of nodes is usually heavily biased towards a few regions of the network, and its statistics play a critical role in the performance seen by users in different places. To the best of our knowledge, we are the first to characterize the age distribution of a dynamic piece of information maintained through gossip in a large system with arbitrary locations, contact rates and input sources. These distributions can be computed as a unique solution of a system of differential equations. This offers both fast numerical estimates and simple linear forms for the asymptotic value of age.

This complete characterization extends recent results on the performance of applications deployed on top of large opportunistic mobile networks [18, 3, 12, 2]. We believe that our result can be significantly extended in several ways: first, by showing that large systems approach the mean-field regime even when their contacts do not follow memoryless statistics. Moreover, since many energy efficient protocols can be modeled as the evolution of nodes in different states, we believe that our method can be applied to understand their performances. Last, in the example we provide (taxicabs in San Francisco), we have observed that a certain number of classes are necessary to describe the diversity of the data sets. We used a simple, regular division of the area of interest into classes. One could study more generally how to best design the number of classes, their shapes and placement in order to minimize the number of classes.

9. ACKNOWLEDGMENTS

The authors would like to thank David McDonald for making his lecture notes available to us. We would also like to gratefully acknowledge Michel Benaïm, and Stratis Ioannidis, for their comments during the preparation of this article. This work was partially funded by the European Commission inside the FET-SAC HAGGLE project (027918).

10. REFERENCES

[1] http://cabspotting.org/.

[2] E. Altman, P. Nain, and J. Bermond. Distributed storage management of evolving files in delay tolerant ad hoc networks. In *Proc. of IEEE INFOCOM*, 2009.

[3] N. Banerjee, M. D. Corner, D. Towsley, and B. N. Levine. Relays, base stations, and meshes: enhancing mobile networks with infrastructure. In *Proc. of ACM MobiCom*, 2008.

[4] C. Bordenave, D. McDonald, and A. Proutiere. A particle system in interaction with a rapidly varying environment: Mean field limits and applications. arXiv:math/0701363v2.

[5] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. MaxProp: Routing for Vehicle-Based Disruption Tolerant Networking. In *Proceedings of IEEE Infocom 2006*, Barcelona, Spain, 2006.

[6] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: spatial mean field regime. Technical report LCA-REPORT-2009-003, EPFL, 2009.

[7] N. Champagnat, R. Ferrière, and S. Méléard. Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. *Theoretical Population Biology*, 69:297–321, 2006.

[8] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *Proc. of ACM PODC*, 1987.

[9] A. Federgruen and H. Groenevelt. The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality. *Oper. Res.*, 34(6):909–918, 1986.

[10] A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proc. of IEEE INFOCOM*, 2005.

[11] IEEE1609.1. IEEE trial-use standard for wireless access in vehicular environments (wave)- resource manager, 2006.

[12] S. Ioannidis, A. Chaintreau, and L. Massoulié. Optimal and Scalable Distribution of Content Updates over a Mobile Social Network. In *Proc. of IEEE INFOCOM*, 2009.

[13] T. G. Kurtz. *Approximation of Population Processes*. SIAM, 1981.

[14] D. McDonald. Lecture Notes on Mean Field Convergence, 2007.

[15] R. Rudnicki and R. Wieczorek. Fragmentation & coagulation models of phytoplankton. *Bull. Pol. Acad. Sci. Math.*, 54:175–191, 2006.

[16] E. Seneta. *Nonnegative Matrices and Markov Chains*. 2nd edition, Springer, 1981.

[17] A. Vahdat and D. Becker. Epidemic routing for partially-connected ad hoc networks. Technical Report CS-2000-06, UCSD, 2000.

[18] X. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance modeling of epidemic routing. *Computer Networks*, 51(10):2867–2891, 2007.

11. PROOFS

11.1 Theorem 4.2

We introduce the following change of variable. Let $\mathcal{A} = \{\ (x,y) \in \mathbb{R}\times\mathbb{R}\ \mid x \geq 0\,,\ x+y \geq 0\ \}$ and for any $c \in \mathcal{C}$ define G_c as the function $G_c : \mathcal{A} \to [0;1]$ s.t $G_c(x,y) = F_c(x, x{+}y)$. Thus, for any $z \geq 0, t \geq 0$, $F_c(z,t) = G_c(z, t{-}z)$ and for any $(x,y) \in \mathcal{A}$:

$$\frac{\partial G_c(x,y)}{\partial x} = \frac{\partial F_c(z,t)}{\partial z}(x, x+y) + \frac{\partial F_c(z,t)}{\partial t}(x, x+y)$$

Let us choose $y \in \mathbb{R}$ arbitrarily, it follows that the function $g_y : x \mapsto G(x,y)$ is a derivable function that satisfies the ODE problem in Lemma 11.1 with $a = 0$, $d' = u(y|d)$ whenever $y \geq 0$ and $a = -y$, $d' = d$ otherwise. It follows from the lemma that $G(x,y) = h(x|0, u(y|d))$ whenever $y \geq 0$ and $G(x,y) = h(x+y|F(-y,0), d)$ otherwise, which, after substitution, proves the final result. We used the following lemma, the proof of which is a straightforward of the Cauchy-Lipschitz theorem.

LEMMA 11.1. *Given $a \in \mathbb{R}$, $b \in [0;1]^C$, $d' \in [0;1]^C$, such that $\sum_i d'_i = 1$, let $v^{(a;d')} : [a;\infty[\to [0;1]^C$ be the solution of the ODE Eq. (2) such that $v^{(a;d')}(a) = d'$. Consider the following ODE problem for a function $H : [0;\infty[\to [0;1]^C$:*

$$\begin{cases} \forall c \in \mathcal{C}\,, & \dfrac{d\ H_c(x)}{dx} = \displaystyle\sum_{c' \neq c} \rho_{c',c} H_{c'}(x) - \left(\sum_{c' \neq c} \rho_{c,c'}\right) H_c(x) \\ & \quad + \left(v^{(a;d')}(x) - H_c(x)\right)\left(\mu_c + \displaystyle\sum_{c'} 2\beta_{\{c,c'\}} H_{c'}(x)\right) \\ \forall c \in \mathcal{C}\,, & H_c(a) = b_c \end{cases} \tag{13}$$

There is a unique solution, given by $H_c(x) = h(x-a|b, d')$.

11.2 Eq. (6)

It follows from standard results on drift and jump processes [?] that $(\mathcal{G}^N\phi)(\nu) = G1 + G2 + G3$ with

$$G1 = \lim_{dt\to 0}\left[\phi\left(\frac{1}{N}\sum_n \delta_{x_n+dt}\right) - \phi(\nu)\right]$$

$$G2 = \mu \sum_{n_0=1}^{N}\left[\phi\left(\frac{1}{N}(\delta_0 - \delta_{x_{n_0}} + \sum_n \delta_{x_n})\right) - \phi(\nu)\right]$$

$$G3 = \frac{2\eta}{N-1}\sum_{m_0<n_0}\left[\phi\left(\frac{1}{N}((\delta_{x_{m_0}} - \delta_{x_{n_0}})1_{\{x_{m_0}<x_{n_0}\}}\right.\right.$$
$$\left.\left. + \quad (\delta_{x_{n_0}} - \delta_{x_{m_0}})1_{\{x_{n_0}<x_{m_0}\}} + \sum_n \delta_{x_n})\right) - \phi(\nu)\right]$$

We now compute $G1$ to $G3$ when $\phi(\nu) = <h,\nu>$.

$$\begin{aligned} & \phi\left(\frac{1}{N}\sum_n \delta_{x_n+dt}\right) - \phi\left(\frac{1}{N}\sum_n \delta_{x_n}\right) \\ = \ & \frac{1}{N}\sum_n (h(x_n+dt) - h(x_n)) \\ = \ & \frac{1}{N}\left(\sum_n h'(x_n)dt + \frac{1}{2}\sum_n h"(y_n)(dt)^2\right) \\ = \ & <h',\nu> dt + (dt)^2 K(dt) \end{aligned}$$

with $|K(dt)| \leq \frac{1}{2}\sup_x |h"(x)|$ (here y_n are some numbers in the range $[x_n, x_n{+}dt]$). It follows that the limit in $G1$ exists, uniformly in t, and $G1 = <h',\nu>$.

$$\begin{aligned} G2 &= \mu\frac{1}{N}\sum_n\left[\sum_{n_0} h(x_n - x_{n_0}1_{\{n_0=n\}}) - h(x_n)\right] \\ &= \mu\,(h(0) - <h,\nu>) \\ G3 &= \frac{2\eta}{N-1}\sum_{m<n}\frac{1}{N}\left[1_{\{x_m<x_n\}}(h(x_m) - h(x_n))\right. \\ &\quad \left. + 1_{\{x_n<x_m\}}(h(x_n) - h(x_m))\right] \\ &= \frac{2\eta}{N-1}\sum_{m\neq n}\frac{1}{N}\left[1_{\{x_m<x_n\}}(h(x_m) - h(x_n))\right] \\ &= \frac{2\eta}{N-1}\sum_{m,n}\frac{1}{N}\left[1_{\{x_m<x_n\}}(h(x_m) - h(x_n))\right] \\ &= \frac{2\eta N}{N-1}\sum_{m,n}\frac{1}{N^2}\left[1_{\{x_m<x_n\}}(h(x_m) - h(x_n))\right] \\ &= \frac{2\eta N}{N-1}\int\int 1_{\{x<y\}}(h(x) - h(y))\,d\nu(x)d\nu(y) \end{aligned}$$

which shows Eq. (6). The proof of Eq. (7) is analogous and is not shown for lack of space.

11.3 Lemma 4.2

Let

$$Z(t) = <h, m(t)> - <h, m_0> - \int_0^t G(h, m(s))ds$$

so that, by hypothesis, $Z(t)$ is a continuous $m(t)-$ martingale. We use similar arguments as in [15, 7]. Apply Itô's formula to the continuous semi-martingale $<h, m(t)>$:

$$<h, m(t)>^2 = <h, m_0>^2 + 2\int_0^t <h, m(s)> G(h, m(s))ds + [Z]_t$$

where $[Z]_t$ is the quadratic variation of Z_t. Compare with the second condition in PROBLEM MP: it follows that $[Z]_t$ is a zero mean $m(t)-$ martingale. By unicity of the Doob-Meyer decomposition, it follows that $[Z]_t = 0$. Thus $Z(t)^2$ is a zero-mean martingale and thus $Z(t) = 0$ a.s.

Now let h span an enumerable convergence determining set H of functions (for example the Fourier basis with rational frequencies). For each $h \in H$ there is a subset E_h, of probability 1, of the probability space Ω over which Equation (9) holds for h. $E = \cup_h E_h$ is also of probability 1 and over E, Equation (9) holds for all $h \in H$. By density, it holds for all h with continuous derivative.

11.4 Lemma 4.3

First we transform Equation (9) into another simpler, but time inhomogeneous problem, as follows. For each $t \in \mathbb{R}^+$, let ϕ_t be the mapping: $\mathbb{R}^+ \to (-\infty, t]$, s.t. $\phi_t(x) = t - x$, and associate to $m(t) \in \mathcal{P}(\mathbb{R}^+)$ the image measure $\ell(t) \in \mathcal{P}((-\infty, t])$ s.t. $\ell(t) = m(t)\circ\phi_t^{-1}$. In particular, for any integrable $h : (-\infty, t] \to \mathbb{R}$,

$$<h, \ell(t)> = <h\circ\phi_t^{-1}, m(t)>$$

If $m(t)$ satisfies Equation (9), then it is differentiable, and for any $h : \mathbb{R}^+ \to \mathbb{R}$ with continuous and bounded derivative:

$$\frac{\partial}{\partial t} <h, m(t)> = G(h, m(t))$$

By standard calculus, it comes that and for any $h : (-\infty, t] \to \mathbb{R}$ with continuous and bounded derivative:

$$\begin{aligned}
\frac{\partial}{\partial t} <h, \ell(t)> &= \frac{\partial}{\partial t} <h \circ \phi_t^{-1}, m(t)> \\
&= <\frac{\partial \phi_t^{-1}(x)}{\partial t} h' \circ \phi_t^{-1}, m(t)> \ + G(h \circ \phi_t^{-1}, m(t)) \\
&= <h' \circ \phi_t^{-1}, m(t)> + <\frac{\partial \phi_t(x)}{\partial x} h' \circ \phi_t^{-1}, m(t)> \\
&\quad +\mu(h(t) - <h, \ell(t)>) + C_1(t, h, \ell(t)) \\
&= \mu(h(t) - <h, \ell(t)>) + C_1(t, h, \ell(t))
\end{aligned}$$

where

$$C_1(t, h, \nu) := 2\eta \int_{-\infty}^{t} \int_{-\infty}^{t} 1_{\{x>y\}}(h(x) - h(y)) d\nu(x) d\nu(y) \tag{14}$$

Thus $\ell(t)$ satisfies the time-inhomogeneous equation:

$$<h, \ell(t)> = <h, \ell_0> + \int_0^t \mu(h(s) - <h, \ell(s)>) + C_1(s, h, \ell(s)) ds \tag{15}$$

for all t and h integrable $(-\infty, t] \to \mathbb{R}$. We now show that, for each initial condition ℓ_0, (15) has at most one deterministic solution $(\ell(t))_{t \in \mathbb{R}^+}$ s.t $\ell(t) \in \mathcal{P}(-\infty, t])$ for all $t \geq 0$.

To this end, let $(\ell(t))_{t \in \mathbb{R}^+}$ and $(\ell'(t))_{t \in \mathbb{R}^+}$ be two solutions with $\ell(0) = \ell'(0)$. Fix $t > 0$ and h differentiable $(-\infty, t] \to \mathbb{R}$. Whenever p is a signed measure on $\mathbb{R}$, we use the notation $|p|$ for its variation and

$$\|p\| = \sup_{h, \|h\|_\infty \leq 1} |<h, p>| = <1, |p|>$$

for its total variation. For $s \in [0, t]$, let $p_s = \ell(s) - \ell'(s)$. Let h differentiable $(-\infty, t] \to \mathbb{R}$, bounded by 1. Then,

$$\begin{aligned}
&\left|C_1(s, h, \ell(s)) - C_1(s, h, \ell'(s))\right| = \\
&\quad \left| 2\eta \int_{-\infty}^{t} \int_{-\infty}^{t} 1_{\{x>y\}}(h(x) - h(y)) d\ell(s)(x) dp(s)(y) \right. \\
&\quad \left. +2\eta \int_{-\infty}^{t} \int_{-\infty}^{t} 1_{\{x>y\}}(h(x) - h(y)) dp(s)(x) d\ell'(s)(y) \right| \\
&\leq \ 2\eta \int_{-\infty}^{t} \int_{-\infty}^{t} 1_{\{x>y\}} \left|(h(x) - h(y))\right| d\ell(s)(x) d\left|p(s)\right|(y) \\
&\quad +2\eta \int_{-\infty}^{t} \int_{-\infty}^{t} 1_{\{x>y\}} \left|(h(x) - h(y))\right| d\left|p(s)\right|(x) d\ell'(s)(y) \\
&\leq \ 8\eta \left\|p(s)\right\|
\end{aligned}$$

Thus, using (15):

$$<h, \ell(t) - \ell'(t)> \leq \int_0^t (\mu + 8\eta) \left\|p(s)\right\| ds$$

and thus

$$\|p(t)\| \leq \int_0^t (\mu + 8\eta) \left\|p(s)\right\| ds$$

and this is true for each $t \geq 0$. By Grönwall's lemma it follows that $\|p(t)\| = 0$ and thus $\ell(t) = \ell'(t)$ for all $t \geq 0$.

It follows easily that two solutions $m(t), m'(t)$ of the problem in Equation (9) must have $\ell(t) = \ell'(t)$ for all $t \geq 0$ and thus $m(t) = m'(t)$ for all $t \geq 0$.

11.5 Lemma 4.4

Integrate by parts in $<h, m(t)> = \int h(z) \frac{\partial F(z,t)}{\partial z} dz$

11.6 Lemma 4.5

We first compute the modulus of continuity of M^{N_k} ([?]):

$$J^{N_k} := \int_0^\infty e^{-u} \sum_j 2^{-j} \left| <f_j, M^{N_k}(u^-)> - <f_j, M^{N_k}(u)> \right| du$$

where f_j is a convergence determining set of bounded functions, uniformly bounded by 1. Since a jump in M^N is almost surely due to a change in the state of one node at a time, it follows that

$$\left| <f_j, M^N(u^-)> - <f_j, M^N(u)> \right| \leq \frac{2}{N} \sup_x |f_j(x)| \leq \frac{2}{N}$$

thus $J_{N_k} \to 0$ almost surely and in distribution. By [?] Theorem 10.2 (a), chapter 2, it follows that M has continuous sample paths.

11.7 Lemma 4.6

From [14], it follows that it is sufficient to show the tightness of the law of X_1^N. From Kurtz [13], Corollary 4.3, it is sufficient to show that for all $\epsilon, T > 0$ there exists a bound $K \in \mathbb{R}^+$ such that

$$\inf_N \mathbb{P}\left(X_1^N(t) \leq K \text{ for all } t \leq T \right) \geq 1 - \epsilon$$

This follows from the fact that

$$X_1^N(t) \leq t - T_1^-(t) \tag{16}$$

where $T_1^-(t)$ is the last time before t when a direct dissemination from the source to node 1 took place. The direct dissemination to node 1 is a Poisson process with rate μ, thus $t - T_1^-(t)$ is exponentially distributed with rate μ and

$$P\left(X_1^N(t) \leq K \right) \geq \mathbb{P}(t - T_1^-(t) \leq K) = 1 - e^{-\mu K} \tag{17}$$

11.8 Eq. (10)

For a single class c, as all nodes belong to this class, we have by definition $u_c(t) = 1$ for any t.

The Cdimensional ODE which defines the solution h used in the expression of F becomes a uni-dimensional real ODE, defined for any function $H : [0; \infty[\mapsto [0; 1]$:

$$\begin{cases} \frac{d\ H(x)}{dx} = (1 - H(x))\,(\mu + 2\eta H(x)) \\ H(0) = b \end{cases} \tag{18}$$

where we denoted $\beta_{\{c\}}$ by η, and all index of class c have been removed as they are not required anymore.

Let $x \mapsto h(x|b)$ be the solution of the above system of equations. The differential equation satisfied by h may be rewritten as

$$h'(x) \left(\frac{1}{1 - h(x)} + \frac{1}{h(x) + \mu/(2\eta)} \right) = \mu + 2\eta \,.$$

By integration this proves that the following function is a constant

$$x \mapsto \ln\left(\frac{h(x) + \mu/(2\eta)}{1 - h(x)} \right) - (\mu + 2\eta)x \,,$$

which proves, for any $x \geq 0$

$$h(x|b) = 1 - \frac{2\eta + \mu}{2\eta + \frac{2\eta b + \mu}{1 - b} e^{(\mu + 2\eta)(x)}} \,.$$

The result then follows from Theorem 4.2.